École Centrale Lyon

MOS 4.4
Nouvelles Technologies de l'Information et de la Communication

Veille sur la NLP

Élève :
Jean Wolff

Enseignant :
Nicolas Jardin

12 mars 2020

Table des matières

1 Introduction

Un ingénieur se doit d'être au courant des avancées technologiques de son domaine, de se forger son opinion et de faire sa part dans le développement de nos sciences et techniques. Or, avec les innombrables laboratoires de recherche à travers le monde, tous les médias et toutes les sources d'information, effectuer un sérieux travail de veille technologique peut être très fastidieux et chronophage. Cette activité importante, parfois délaissée aux stagiaires dans certaines entreprises, peut devenir très efficace si l'on utilise des outils adaptés, et apporte alors beaucoup de valeur.

Dans le cadre du MOS 4.4 "Nouvelles Technologies de l'Information et de la Communication" de l'Ecole Centrale de Lyon, nous développons des outils de veille pertinents par rapport au domaine qui nous intéresse. Le domaine que j'ai choisi est le traitement du langage naturel en informatique, plus fréquemment appelé « natural language processing » (NLP).

La NLP est une partie phare des technologies de l'information : à la jonction de la linguistique, de l'informatique et de l'intelligence artificielle, c'est la partie de cette dernière qui a le plus d'applications. Les principales sont les moteurs de recherche, tels Google et Bing, des chatbot de plus en plus « intelligents », l'analyse de sentiment, la création de résumés... L'application historique de l'apprentissage artificiel appliqué au langage est la détection des spams, pour lesquelles gmail réussi depuis 2004 à bien classer plus de 99% des e-mails.

2 Sources surveillées

La veille sera appliquée à 4 types de sources d'informations :
- Les entreprises innovantes majeures
- Les publications de ressources de Google Scholar
- La presse spécialisée
- Les réseaux sociaux

La quasi-totalité des sources repérées sont en anglais.

Les entreprises qui publient les papiers de recherche les plus pertinents de la NLP sont Google et Facebook :
- https://ai.facebook.com/
- https://www.blog.google/

Google Scholar est la ressource majeure de publications scientifiques pour tout le domaine de l'intelligence artificiel.
- https://scholar.google.fr/

On peut trouver d'innombrables sites web proposant des cours en ligne, des blogs reprenant des articles scientifiques, et des articles apportant des réflexions très intéressantes. Il serait vain de tenter d'en établer une liste exhaustive, d'autant qu'elle varierait dans

le temps. Cela souligne l'importance d'adopter des outils de veille efficaces. Les médias consacrés à la NLP apportant les informations les plus pertinentes sont :

— https://medium.com/
— https://www.theregister.co.uk/

Enfin, les réseaux sociaux permettent d'obtenir des informations d'une manière plus horizontale et décentralisée :

— https://twitter.com/home nlp
— https://www.reddit.com /r/textdatamining

3 Mots-clés et requêtes effectuées

Il est certain que l'ensemble de ces ressources tous les développements importants du domaine étudié. Il faut toutefois définir avec précision les mots-clés et les recherches effectuées, afin d'effectivement extraire toutes les informations intéressantes.

Le domaine du traitement du langage naturel possède un acronyme clair : « NLP ». Ces 3 lettres sont systématiquement définies parmi les mots-clés de ceux qui émettent des publications, francophones ou anglophones, ce qui permet d'obtenir facilement, pour chacune des sources, les informations spécifiquement en rapport avec le domaine étudié.

De nombreuses autres recherches ont été effectuées, sans apporter davantage de résultats pertinents.

4 Outils de collecte d'information

4.1 Agrégation de flux RSS

Pour ne pas avoir à effectuer l'activité chronophage de consulter chaque ressource une à une, il est indispensable d'utiliser un agrégateur de flux RSS. Celui utilisé est Feedly, mais la version payante semble malheureusement nécessaire pour pouvoir partager la collection créée. Toutefois, elle est disponible à l'adresse https://feedly.com/, avec les identifiants suivants :

— login : jean.wolff@ecl16.ec-lyon.fr
— mot de passe : veilleNTIC

4.2 Création de flux RSS

Certains sites ne proposent pas de flux RSS (ai.facebook, blog.google, scholar.google), alors il faut les créer, par exemple avec FeedFry. Néanmoins, FeedFry ne permet pas toujours d'extraire le flux RSS associé à une requête. Par exemple pour ai.facebook, on récupère toutes les publications du site, et non celles obtenues en recherchant « NLP » dans leur moteur de recherche. Pour éviter le noyage d'information, on peut créer une collection à part sur Feedly pour ces sources dont l'extraction du flux RSS posent problème.

Par ailleurs, la version payante de Feedly est nécessaire pour importer des flux RSS de Twitter. On pallie à ce problème avec Feedfry, qui convient très bien.

5 Outils de curation

Pour les 3 étapes de la curation, valider les articles pertinents, les commenter puis les organiser, l'outil Wakelet a été utilisé, pour sa simplicité de prise en main et sa clarté. Le Wakelet créé est disponible à l'adresse hhttps://wke.lt/w/s/oLAFoZ. Pour une raison incomprise, il est toutefois impossible de légender et commenter les liens issus de certains site.

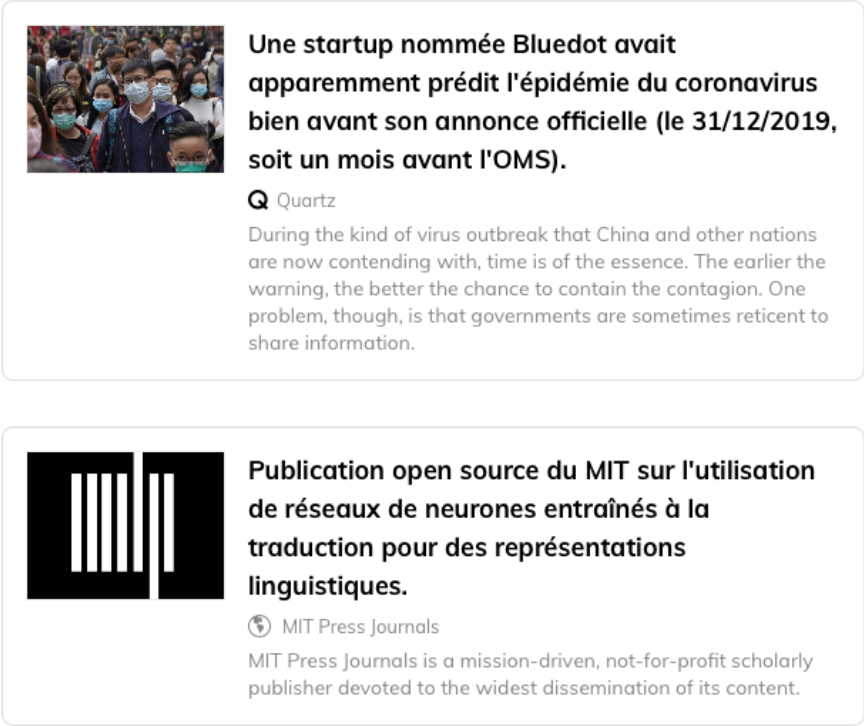

FIGURE 1 – Visualisation de deux éléments du Wakelet.

Le livrable produit à l'issue de cette veille sur la NLP est consultable à l'adresse https://jeanwolff10.github.io/VeilleNTIC_NLP/.

6 Conclusion

Le dispositif de veille mis en place est représenté sur le schéma ci-dessous.

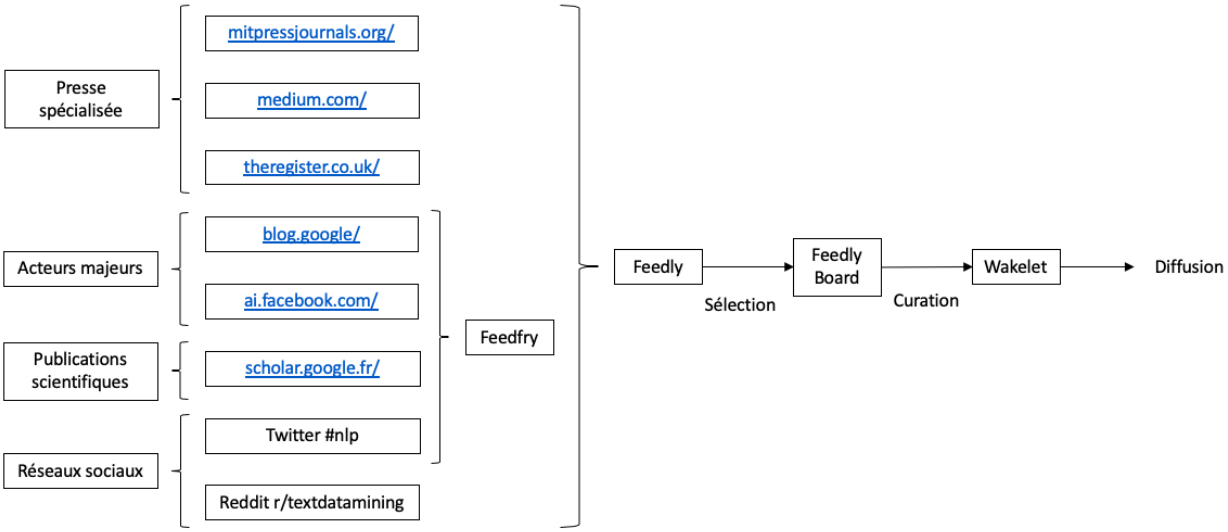

FIGURE 2 – Schéma du dispositif de veille mis en place

Ce dispositif, dorénavant bien paramétré et fonctionnel, permet de s'informer efficacement sur les dernières actualités du domaine de la NLP. La mise au point de cette veille a demandé du temps, mais cela sera largement compensé par le gain de temps considérable qui sera réalisé à l'avenir, à chaque fois que je voudrais mettre à jour mes connaissances.